Received: 2 November 2018 | Revised: 8 October 2019 | Accepted: 8 November 2019

DOI: 10.1002/tea.21610

RESEARCH ARTICLE

JRST | WILEY

Evaluating the evidence base for evidence-based instructional practices in chemistry through meta-analysis

Md Tawabur Rahman | **Scott E. Lewis**

Department of Chemistry, University of South Florida, Tampa, Florida

Correspondence
Scott E. Lewis, Department of Chemistry, University of South Florida, Tampa, FL.
Email: slewis@usf.edu

Funding information
National Science Foundation, Grant/Award Number: DUE-1712164

Abstract

Investigating the effectiveness of instructional practices provides an evidence base to inform instructional decisions. Synthesizing research studies on instructional effectiveness provides an estimate of the generalizability of effectiveness across settings, along with an exploration of factors that may moderate the impact, which cannot be achieved within individual studies. This study sought to provide a synthesis of evidence-based instructional practices (EBIPs) particular to chemistry through meta-analysis. Ninety-nine studies were analyzed comprising a broader view of chemistry specific studies than past meta-analyses. The results showed that EBIPs feature a demonstrably positive impact on students' academic performance in chemistry, although assessment topic coverage and setting size emerged as relevant moderators of impact and prevented making definitive conclusions of the relative impact of each EBIP. In examining publication bias, an asymmetric distribution of studies based on standard error *(SE)* and effect size was found, indicative of potential publication bias. To explore the potential impact of bias, the trim and fill method was employed resulting in a range for the overall weighted effect size from 0.29 to 0.62. The study concludes that evidence-based instructional practices have demonstrated effectiveness even in consideration of potential publication

bias, as the range of effect sizes remains positive, but highlights the continued need to publish null findings in the research literature.

KEYWORDS

chemistry, effect size, evidence-based instructional practices, cooperative learning

1 | INTRODUCTION

Meta-analyses in education research offer unique insights into a research field via the synthesis of research literature, yet efforts to conduct these analyses have largely relied on grouping literature based on general designations (e.g., science or math). This technique provides limited information on what is known on discipline-specific instructional practices (e.g., biology or chemistry) in secondary or postsecondary education. This study synthesizes the research literature on the effectiveness of evidence-based instructional practices (EBIPs) from a discipline-specific perspective. In postsecondary chemistry, several different EBIPs have propagated through nationally disseminated initiatives, each advancing a particular variant of active learning. This study uses meta-analysis to explore the evidence base for several EBIPs in chemistry by investigating the relative effectiveness among EBIPs, the factors that may explain variation in effectiveness and the extent publication bias may modify the reported effectiveness.

1.1 | Meta-analyses in science and chemistry education

Two recent meta-analyses investigating the impact of instructional pedagogies on students' academic performance in science have been conducted. Ruiz-Primo, Briggs, Iverson, Talbot, and Shepard (2011) searched 27 journals determined by an interdisciplinary advisory board and article recommendations by the same board to locate articles evaluating instructional interventions in science and engineering at the postsecondary level. The journals were searched for terms related to active learning, inquiry and problem-based learning with search terms varying based on the journal. Their search identified 166 studies that fit their selection criteria and of those studies 20 were conducted in chemistry. The weighted overall effect size, the difference between group means divided by the standard deviation *(SD)*, observed was 0.50 and for the 20 chemistry articles was 0.46.

Freeman et al. (2014) reviewed all articles in 55 journals, searched seven databases including Web of Science, PubMed, ERIC and ProQuest, reviewed past meta-analyses and the references for all identified studies (snowball sample) for articles evaluating instructional interventions in STEM education at the postsecondary level. The databases and journals were searched for terms related to audience response system (clickers), cooperative learning, collaborative learning, case-based learning, problem-based learning, peer instruction and workshops with search terms varying based on the database. The search resulted in 225 identified studies of which 22 were in chemistry. The overall weighted effect size observed was 0.47 and the effect size for chemistry was approximately 0.40. While both meta-analyses synthesize a substantive

database of education research articles, each offer a notably smaller number of studies related to chemistry.

Meta-analyses particular to chemistry have also been conducted but feature comparable numbers of chemistry studies to the aforementioned studies. Warfa (2016) searched seven journals and five databases for the keywords cooperative learning paired with chemistry. The search resulted in 25 articles and an average weighted effect size, measured by Hedges' *g*, of 0.68. Apugliese and Lewis (2017) conducted a follow-up study on the corpus of studies identified by Warfa (2016), including an adjustment for pretests and found a weighted average effect size of 0.59. Leontyev, Chase, Pulos, and Varma-Nelson (2017) identified chemistry articles from a review article on Peer-Led Team Learning and located 16 studies with an average weighted effect size, measured by Hedges' *g*, of 0.37. Each chemistry specific meta-analysis investigates a single EBIP (e.g., Peer-Led Team Learning) and as a result each analyses 25 or fewer studies; this number of studies is comparable to the number of chemistry specific studies analyzed in meta-analyses on the broader fields of science or STEM education. Other meta-analyses have been conducted on POGIL (Walker & Warfa, 2017) and blended learning, combining face-to-face instruction with computer mediated instruction (Vo, Zhu, & Diep, 2017), but these analyses were not subject specific. Thus, the current literature is unable to provide a thorough synthesis of research on effective instructional practices particular to chemistry or evaluate the effectiveness of a particular EBIP in chemistry relative to other widely used EBIPs.

1.2 | Publication bias in science education and chemistry education

A unique advantage of meta-analyses is the ability to examine trends among published studies that may be indicative of publication bias. Publication bias is the phenomenon where studies that exhibit significant effect sizes are more likely to be submitted and/or accepted to peer-reviewed journals than studies with null or negative effect size. The presence of publication bias has the potential to overstate the overall effect (Becker, Rothstein, Sutton, & Borenstein, 2005). In such a case, the interpretation of the effectiveness of treatment over control will be misleading as the true effect is lower due to the presence of publication bias in a meta-analysis. Among the recent meta-analyses in science education, Freeman et al. (2014) conducted the following tests: inspection of a funnel plot, rank correlation test, Egger's regression test, fail-safe N and a trim and fill method. On studies investigating student assessment outcomes they found significant relationships between *SE* and effect size, an Orwin's fail-safe N value of 114 studies with null results to move the overall effect size down to a small effect, and that trim and fill found a consistent effect size of 0.47 (confidence interval 0.37–0.56). The authors concluded there was no indication that publication bias influenced their results.

In chemistry specific meta-analyses, Warfa (2016) found a significant intercept for Egger's regression test and a nonsignificant value for the rank correlation test. A visual inspection of the funnel plot found higher effect sizes with smaller sample sizes. The Orwin's fail-safe N was 23 studies for the overall effect size to reach nonsignificance and that trim and fill maintained the effect size at 0.68 (confidence interval 0.34–0.83). Warfa (2016) concluded that any presence of publication bias within the corpus of identified studies was not likely to alter the overall conclusions. Leontyev et al. (2017) conducted a trim and fill analysis on their database and did not report the updated effect size but indicated that it did not reveal substantial publication bias. They cautioned against reliance on this finding owing to high variation and a small number of

studies. In summary, the studies presented show minimal evidence of publication bias within science education or chemistry education studies. However, the lack of a sizable corpus of chemistry studies included in any one analysis prevents a strong conclusion regarding the presence of publication bias particular to chemistry education research.

1.3 | Rationale

Past efforts to synthesize educational research in chemistry can be found either within a large corpus of studies in meta-analyses conducted on science education or STEM education or in narrowly defined chemistry meta-analyses. Both approaches have generated a small corpus of chemistry specific studies, with the largest analysis having 25 studies. Of the meta-analyses that span multiple disciplines in STEM (Freeman et al., 2014; Ruiz-Primo et al., 2011) the results show considerable variation of effectiveness by discipline as shown in Table 1. This volatility by discipline calls to question the extent that the overall, combined results across disciplines are applicable to a specific discipline and leads to the possibility that a discipline-specific meta-analysis would generate unique results. Past meta-analyses including those across disciplines and those specific to chemistry explored a single or small set of search words. By incorporating a set of search terms targeting a range of instructional practices it is possible to generate a more comprehensive synthesis of chemistry education literature than previously done. Creating a set of search terms requires a discipline-specific perspective, as instructional practices highly visible within one discipline are not as well known in other disciplines. By generating such a meta-analysis instructors and researchers would be informed by the current evidence base for a variety of instructional practices tested within a chemistry instructional setting.

Further, by analyzing a sizable corpus of chemistry education research articles it is possible to make a substantive investigation of potential publication bias within chemistry education. Past meta-analyses that have investigated publication bias across multiple disciplines may lack sensitivity to such bias within a particular discipline. Publication bias can result from the viewpoints of authors, reviewers, and editors decisions made when presented with null or negative results. We argue that these decisions are likely discipline-specific as chemistry education represents a research culture where chemistry education researchers often submit to chemistry education journals and are reviewed by other chemistry education researchers. In line with this

TABLE 1 Past meta-analyses demarcated by discipline

	Freeman et al. (2014)			Ruiz-Primo et al. (2011)		
Subject	***k***	**Hedges' *g***	***SE***	***k***	**Hedges' *g***	***SE***
Biology	33	0.30	0.11	53	0.45	0.08
Chemistry	22	0.39	0.14	20	0.46	0.07
Computer science	8	0.31	0.25			
Engineering	19	0.48	0.15	22	0.11	0.11
Geology	2	0.52	0.49			
Mathematics	29	0.34	0.12			
Physics	31	0.72	0.11	71	0.58	0.04
Psychology	14	0.61	0.15			

position, publication bias from a discipline-specific perspective would serve to inform that discipline and highlight the importance of the phenomenon to other disciplines.

This study aims to address these research gaps regarding synthesizing the research literature on several EBIPs within chemistry. In so doing this study will characterize the effectiveness of each EBIP particular to chemistry and in particular will facilitate an exploration into instructional characteristics that moderate effectiveness and characterize the limitations in generalizability of the current state of research. Additionally, by considering multiple EBIPs this study allows the possibility to characterize the research base and effectiveness of each EBIP relative to other EBIPs. The results from this analysis can then serve to inform instructors about the current state of research literature on effective instructional practice in chemistry and inform chemistry education researchers about areas where future research is needed. This study will also examine evidence of potential publication bias, which is necessary to understand the impact this bias may have on the reported effectiveness in chemistry education research. As a result, this study will pursue the following research questions:

1. What is the evidence base on the effectiveness for several evidence-based instructional practices on student academic performance in chemistry?
2. What is the relative effectiveness of each evidence-based instructional practice relative to other widely studied practices in chemistry?
3. What is the evidence that publication bias may be present in evaluating EBIPs in chemistry? With sufficient evidence for bias, what impact would it have on interpreting the above findings?

1.4 | Evidence-based instructional practices

Stains and Vickrey (2017) characterize EBIPs as instructional practices designed to improve student academic performance that are developed and supported by a research base that investigates the impact of the practice on student academic performance. EBIPs describe a wide range of instructional practices in chemistry and no exhaustive list of EBIPs in the literature is available. As a result, it is not possible to characterize the evidence base for all EBIPs. Instead this review focused on a subset of EBIPs in chemistry selected from their inclusion in recent reviews of chemistry education research (Eberlein et al., 2008; Seery, 2015; Warfa, 2016): Process-Oriented Guided Inquiry Learning, Peer-Led Team Learning, Problem-Based Learning, cooperative learning, collaborative learning, and flipped instruction. Additional instructional practices including, but not limited to, the science-writing heuristic, argument-driven inquiry, writing-to-learn and the incorporation of animations, have substantive evidence bases but are not included herein owing to the scope of the study. Additionally, the nature of a meta-analytical approach requires combining evidence bases that arise from similar research designs. This investigation focuses on quasi-experimental and experimental comparisons given their frequency in the research literature (Mack, Hensen, & Barbera, 2019). Other investigative approaches such as qualitative investigations into the quality of students' written responses or quantitative measures of growth over time, generate compelling evidence in support of instructional practices but cannot be synthesized with a corpus of studies enacting comparative designs. As a result, the scope of the current study is limited to characterizing the evidence base for the subset of EBIPs described and only including evidence generated from quasi-experimental and experimental comparisons. To better

characterize the selected EBIPs a brief description and an example instructional practice for each EBIP follows.

1.4.1 | Cooperative learning

Cooperative learning is a general term used to describe students working together on a common task. Johnson, Johnson, and Smith (1998) describe essential features of effective cooperative learning as positive interdependence, accountability, promotive interactions, teaching interpersonal skills and group processing. Positive interdependence describes a perception that each member's contribution will benefit all members of the group. Accountability requires that the group and each individual be assessed and provided meaningful feedback and if needed additional resources. Promotive interactions require regular communication among group members that serve to encourage each member and reaffirm the commitment made by each member of the group. Teaching interpersonal skills is an explicit incorporation by the instructor in modeling how to engage in a team. Finally, group processing describes a reflective aspect where the group self-evaluates its progress and adapts as necessary.

An instructional example of cooperative learning in chemistry could involve the teaching of chemical kinetics, also termed reaction rates. An instructor using this technique may assign students to groups and provide a series of problems for the group to work on. As part of the instructional technique, the instructor may model productive behavior in the group or provide feedback to students on their contributions to the group. Assessments may include assessing students individually upon the completion of cooperative learning, assessing the group on their performance on the task or including a component that evaluates students' contributions to the group

1.4.2 | Collaborative learning

Collaborative learning shares much in common with cooperative learning in that both rely on group work but is differentiated by collaborative learning emphasizing students creating knowledge through social interactions (Barkley, Cross, & Major, 2014). In one example of collaborative learning students are placed within a group with a common objective to learn a concept or skill. The concept or skill is broken down into subcomponents and each member of the group is assigned one subcomponent to learn. When the group reconvenes, each member is responsible for presenting their subcomponent to the group so that each group member becomes familiar with the entire concept or skill. In one variant of collaborative learning, termed jigsaw, a member assigned a particular subcomponent meets with members from the other groups in the class assigned the same subcomponent, thus creating a secondary group focusing on a particular subcomponent. In jigsaw, the original group still reforms as in collaborative learning to present the subcomponent to the original group members.

As an example of teaching chemical kinetics with collaborative learning, students within a group could be assigned a subcomponent to explore the impact of concentration of each reactant, temperature and the presence of a catalyst on the reaction rate. This exploration could include a lab component where these parameters are physically manipulated, a review of experimental evidence presented to the students or a review of reference literature. The students would then present each subcomponent to the original group to build a comprehensive picture

of the factors that influence reaction rates. In a jigsaw variation, the process would be the same but each subcomponent investigation would happen in groups; for example, each student tasked with exploring the impact of temperature would work together to conduct this exploration.

1.4.3 | Problem-based learning (PBL)

PBL instruction places students in groups working on a contextually framed problem (Eberlein et al., 2008; Gijbels, Dochy, Van den Bossche, & Segers, 2005). Students are tasked to create a process to identify the information needed to address the problem, enact the process to collect the information and propose a solution to the problem. The procedure may be iterative where gaining information leads to refining the planned process for addressing the problem. Finally, students generate a proposed solution to the contextual problem.

An instructional example of teaching chemical kinetics with problem-based learning may be to provide students the task of maximizing the rate of a chemical reaction in the context of a chemical industry setting with a cost-basis framework. Students would be directed to make a plan on how to gather the needed information on the chemical reaction, enact the plan and if necessary repeat the process until they develop a proposed solution to the problem.

1.4.4 | Process oriented guided inquiry learning (POGIL)

POGIL is a small group, lecture-free instructional method (Minderhout & Loertscher, 2007) with two distinct components: process skills and guided inquiry. Process skills include communication skills, teamwork, problem solving, critical thinking, group management, information processing and self-assessment ("Process oriented guided inquiry learning," 2018). To facilitate process skills, students are assigned particular roles within their group such as manager, reflector, and presenter (Farrell, Moog, & Spencer, 1999). To ensure each student gains experience with the range of skills, assigned roles are often rotated among students within a group ("Process oriented guided inquiry learning," 2018). Student groups practice process skills while engaging in guided inquiry (Eberlein et al., 2008). Guided inquiry follows a three-phase learning process: first is the exploration phase where students develop the desired content from the model provided to them; second is concept development where students learn about new terminology and/or links between the prior knowledge and the newly developed concept; and finally students apply the concept to new situations to demonstrate the utility of the newly learned concept.

As an example of POGIL designed to teach chemical kinetics, students would be assigned a small group and each student would be assigned a role to carry out throughout the activity. The group would be provided a series of experimental data from POGIL instructional materials or from a laboratory experiment regarding the rate of a chemical reaction. The group would be provided a series of questions that prompts the group to analyze the data provided and construct a mathematical model of the rate law. Upon creation of the mathematical model, the group would be introduced to the terminology and components of a rate law. Finally, the group would be tasked with applying the developed rate law to additional situations or explore the utility of other rate laws and presenting their findings to the rest of the class.

1.4.5 | Peer-led team learning (PLTL)

PLTL relies on peer leaders, students who succeeded in a target course, returning to that target course to lead small groups of students in a session called a workshop. Designers of PLTL describe six critical components of the pedagogical approach: (a) workshops are integral to the course, (b) instructors are involved in selecting materials and training and supervising peer leaders, (c) peer leaders are trained and supervised, (d) workshop materials are appropriately challenging and related to course content, (e) workshops are 2 hr per week with students working in groups of six to eight, and (f) institutional support for the adoption (Wilson & Varma-Nelson, 2016).

In teaching chemical kinetics using PLTL, students may first attend lecture or in-class activities that present chemical kinetics. Instructors would then design workshop materials related to chemical kinetics and train peer leaders on those materials. The training would attempt to model the workshop session by instructors challenging peer leaders with different scenarios that students may encounter. Students would then meet with their peer leader in the workshop and work as a group on the materials. In this setting, the peer leader's primary responsibility is to facilitate group work by serving as a resource when the group is stuck and challenging the group to ensure all group members are involved and all members can explain the group's consensus.

1.4.6 | Flipped classes

The flipped class approach involves presenting content outside of class to facilitate active learning within the class. The presentation of content frequently includes instructional videos, which can be created by the instructor or identified among existing resources, but can also take the form of assigned readings (Seery, 2015). The movement of content to outside the formal class meeting environment allows for class meeting time to be dedicated to active learning. Active learning can take a wide variety of forms and can include students' discussing the content, engaging in a problem set or experiential learning or working in groups employing any of the EBIPs previously discussed (Robert, Lewis, Oueini, & Mapugay, 2016).

In the chemical kinetics example, a flipped class may assign students to watch a small set of instructor created videos on the factors that relate to reaction rates. Then, students may be tasked with an online quiz on the same videos to promote attention to the videos. Finally, in-class, students could work in groups determining rate laws from experimental evidence and using their knowledge of rate laws to make predictions on factors related to reaction rates.

2 | METHOD

2.1 | Criteria for inclusion

To be considered for inclusion in this meta-analysis, each study had to describe an investigation that met the following criteria:

1. An investigation of the effects of an EBIP instructional strategy in a chemistry class.

2. The use of a quasi-experimental or experimental research design where a group of learners that experienced an EBIP pedagogy (experimental) were compared against a reference group (control).
3. The incorporation of a measure of student academic performance in chemistry common to both groups.
4. Sufficient information on student-level data to determine an effect size. Sufficient information includes mean, *SD*, and sample size for each group or inferential statistics such as *t* test or *F* test results with sample size.
5. Published between 2000 and 2017 and reported in English.

2.2 | Article identification

The review and integration of research literature began with the identification of the relevant studies. Web-based searches were conducted on the databases ProQuest, Web of Science, and Scopus and a separate search was conducted of the *ACS (American Chemical Society) Symposium Series* as a repository of chemistry specific work that is not indexed by the databases. These databases were chosen as Web of Science indexes the major journals in chemistry education and science education, ProQuest indexes graduate student dissertations, Scopus indexes journals and dissertations in education research and *ACS Symposium Series* offers an alternative peer-reviewed outlet for chemistry education research. Each database was searched with 16 key phrases: cooperative learning, collaborative, group learning, group work, jigsaw, small groups, student team, team based learning, peer led team learning, peer learning, PLTL, process oriented guided inquiry, process-oriented guided inquiry, POGIL, problem based learning, and flipped. Each key phrase was coupled with "chemistry." Key phrases encompassing more than one word were entered as a phrase within quotes, for example "cooperative learning."

In Scopus each key phrase was searched within the abstract field and chemistry was searched in all fields; in Pro-Quest each key phrase was searched within the abstract field and chemistry was searched in the anywhere field; and in Web of Science both the key phrase and chemistry were searched in the topic field. The set of 64 searches (16 key phrases in each search engine and the symposium series) resulted in 8,325 hits. The following preliminary screenings were performed to identify hits to remove: duplicate hits within the search results, studies from journals (Chimia, chemosphere, chemphyschem, etc.) that do not publish educational research and conference abstracts without an accompanying published text (e.g., American Chemical Society National Meeting presentations). Next study titles were reviewed to identify and remove hits that were clearly unrelated to chemistry education (e.g., engineering education or medical studies) or hits that were secondary reports of the primary literature. Finally, the first author downloaded each publication in case of confusion to check whether those particular publications met the criteria. This review was necessarily conservative, if there was a possibility of inclusion; the article was kept for further analysis. These procedures resulted in a revised total of 702 studies. The researchers found at this level of screening that they needed to further operationalize the first criteria, namely what constituted a chemistry class. The decision was made to include sources pertaining to applied forms of chemistry education (e.g., biochemistry, medical chemistry, and physical chemistry) but studies concerning related fields (e.g., medical students or pharmacy students learning a range of content where chemistry was one part) were removed. This pass resulted in 302 studies that met the stated criteria. The review process is summarized in the PRISMA flow diagram (Moher, Liberati, Tetzlaff, & Altman, 2009) presented in Figure 1.

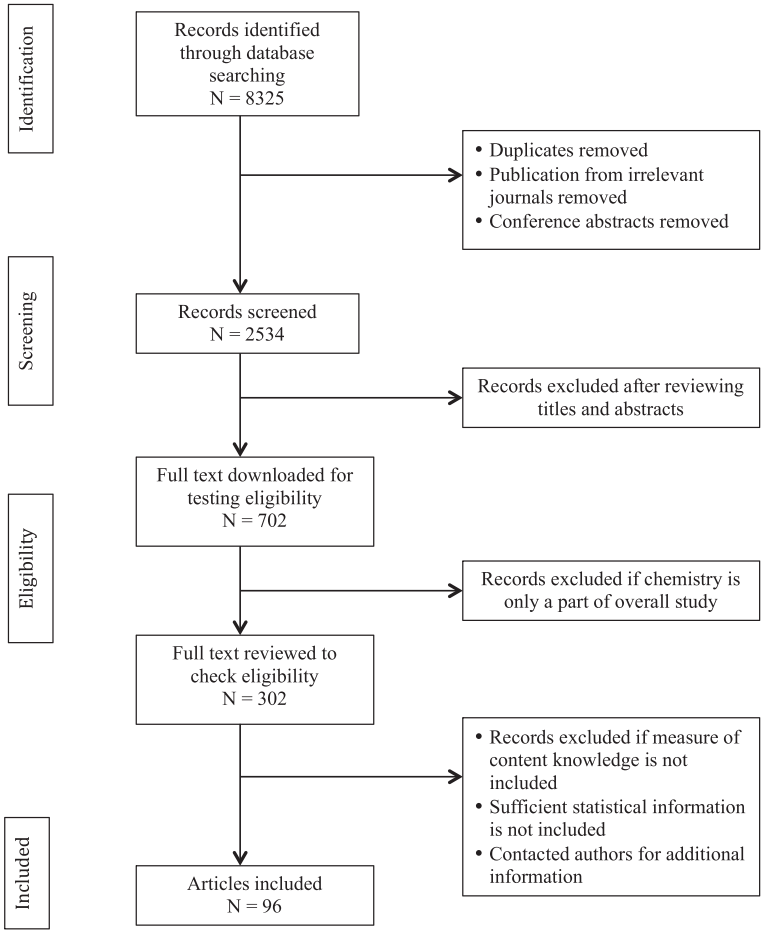

FIGURE 1 PRISMA flow diagram

2.3 | Coding of articles

The 302 identified studies were reviewed for the necessary data described in criteria three and four above. Only 93 of the 302 identified studies contained sufficient information to determine effect sizes. Many studies reported group sizes and average exam scores, but did not report *SD*s. An email and follow-up email was sent out to the corresponding authors of the studies with insufficient information to determine effect size with the response resulting in information for three additional studies to consider for inclusion. Some studies reported multiple tests or multiple semesters of data. In these cases, the decision was made to condense each set of data into one effect size per study using a procedure detailed below. A few studies evaluated two EBIPs independently in comparison to traditional instruction: Ding and Harskamp (2011) compared peer instruction to individual learning and collaborative learning to individual learning, in similar fashion Rau, Kennedy, Oxtoby, Bollom, and Moore (2017) evaluated flipped instruction and collaborative learning, and Doymus, Karacop, and Simsek (2010) evaluated group investigation and collaborative learning. As each EBIP comprised a unique group of students the decision was made to treat each article as representing two distinct studies and calculate two effect sizes. Combined, the inclusion of the data received via email and the decision to

report two effect sizes from the aforementioned three articles meant that the corpus of studies analyzed comprise 99 studies from 96 unique sources.

The 99 datasets were reviewed and coded based on the type of EBIP using the following possible codes: Collaborative, POGIL, PLTL, PBL, Flipped, and nonspecified cooperative learning. The nonspecified cooperative learning represented articles where students worked in groups but no further information was provided that could characterize any of the other EBIPs. One study, Lewis and Lewis (2008) used a combination of PLTL and POGIL in the treatment group, this study was coded as split EBIP use and was treated as undefined EBIP use when analyzing this moderator. Studies were also coded based on the coverage of content within the assessment used owing to past research on the relevance of this construct to moderate effect size (Apugliese & Lewis, 2017). The coding options for content coverage were *cumulative*, measuring student performance on an entire term or semester of content commonly occurring as a final exam, versus *single-topic*, measuring student performance on a defined portion of content in the course commonly occurring as an in-term exam or a topic-specific concept inventory.

2.4 | Calculating effect sizes

To characterize the difference between two groups Cohen's d was calculated, as the difference in means divided by the pooled SD (Lipsey & Wilson, 2001), and then converted to Hedges' g to correct small sample size bias. Hedges' g and SE for each study were calculated using the formulas (Lipsey & Wilson, 2001, p. 72):

$$g = d\left(1 - \frac{3}{4(n_t + n_c) - 9}\right)$$

$$SE = \sqrt{\frac{n_t + n_c}{n_t n_c} + \frac{g^2}{2(n_t + n_c)}}$$

where n_t and n_c are the sample size for the treatment and control respectively. A random-effects model was estimated using the metafor program (Viechtbauer, 2010). Tau-squared was estimated using the DerSimonian and Laird (DL) estimator. The effects of moderators were examined using a mixed-effects model (with moderators fixed and studies random) using metafor and specifying the same random-effects variance estimator.

2.4.1 | Articles with multiple comparisons

To obtain a single effect size data point from multiple comparisons within a single study one of two approaches is followed. For studies that conducted multiple comparisons using the same sample, for example considering a set of examinations across a term, (e.g., Doymus, 2007), a Hedges' g was calculated for each comparison and then averaged to obtain a single effect size for the study. For studies that conducted multiple comparisons with differing sample sizes, for example a study incorporates data from multiple years with the same intervention, (e.g., Baepler, Walker, & Driessen, 2014) a weighted average approach was used. The weighted average exam score for each group (experimental and control) was calculated by multiplying each exam score

by the associated sample size, summing the resulting products, and dividing the sum by the total sample size. Pooled *SD* was computed using the *SD*s provided. Finally, Cohen's *d* was calculated from the weighted average for each group and the pooled *SD* and then converted to Hedges' *g*. Other studies with unique designs such as multiple experimental or control groups or conducting the comparison in different courses (e.g., Casadonte, 2016; Kırık & Boz, 2012; Stoica, Chiru, & Chiru, 2012) were also treated with the weighted approach to generate a single effect size. To investigate the sensitivity of the presented results to the decision to combine dependent data into a single data point, the analyses were also conducted using a robust variance estimate procedure (Tanner-Smith, Tipton, & Polanin, 2016). The results from this procedure are included in the supplemental materials with no substantive changes to the results presented herein.

For studies that used a pretest/posttest design, where the same test was used before and after the instructional intervention (e.g., Özden, 2009), a Hedges' *g* value was calculated for both the pretest and posttest separately and then the value for the pretest was subtracted from the value for the posttest. In studies that used differing tests before and after the intervention, where the items were not identical between administrations, the posttest was used to determine the effect size and the earlier test was not used in determining effect size.

2.5 | Reliability in calculations and coding

Due to nature of the complexity of effect size calculation, particularly in studies with multiple comparisons, each author coded and calculated effect sizes for a set of 20 studies independently. The authors compared the codes and effect size calculated, discussed discrepancies and revised the coding scheme and effect size calculation decisions to clarify the decision making process. This process was continued iteratively on a different set of 20 articles until no further revisions to the coding scheme were made. Finally, a set of 10 studies was coded and effect size calculated resulting in complete agreement between the two authors. The first author coded and calculated the effect sizes for the remaining 49 studies.

For coding of EBIP pedagogy, a study had to refer directly to the name or the acronym for POGIL (process-oriented guided inquiry learning), PLTL (peer-led team learning), PBL (problem-based learning), and flipped instruction. The collaborative code was reserved for studies where students in groups had differentiated tasks. Studies using a jigsaw approach were labeled as collaborative as well. If the study used group work but did not fit the above terms it was coded as nonspecified cooperative learning. For assessment coverage, single-topic had to have a clearly defined topic or small set of topics such as an interim exam that covered two topics or chapters of content. The cumulative assessment code was reserved for an assessment that measured content spanning an entire term (semester or quarter) or longer.

2.6 | Outliers

Studies with extreme effect sizes can disproportionately impact the overall effect in a meaningful way. Each study was characterized based on their effect size relative to the overall average effect size of the entire corpus. Studies that were more than two *SD*s from the overall average effect size were considered outliers and removed from future analyses (Lipsey & Wilson, 2001, p. 108). To explore the impact of this decision, all analyses were repeated with a more

conservative definition of outliers, removing studies more than three *SD*s removed from the overall average, and with retaining all studies.

2.7 | Analyzing publication bias

To explore publication bias among the corpus of studies a visual inspection of funnel plots and statistical tests via rank correlation test and Egger's regression test were conducted. The funnel plot charts *SE* versus effect size so that the top of the plot has small *SE*, associated with larger sample sizes, and the bottom of the plot larger *SE*s. An unbiased data set is expected to have a narrow range of effect sizes at the top, where the *SE* is small, and moving downward on the plot the range of effect sizes should increase symmetrically as *SE* increases. Departures from a symmetrical increase in the range could be interpreted as evidence of bias, as it is indicative that studies with smaller sample sizes had differing effect sizes than larger sample sizes. Rank correlation test and Egger's regression test were used to measure asymmetry from information provided by the funnel plot (Begg & Mazumdar, 1994; Egger, Smith, Schneider, & Minder, 1997). Each test estimates the association between effect size and *SE* with an unbiased data set resulting in a correlation or regression coefficient proximate to zero. The null hypothesis of the coefficient equal to zero can be tested statistically; finding statistical significance leads to rejecting the null hypothesis and supporting the alternative hypothesis of a relationship between *SE* and effect size, seen as evidence of an asymmetric distribution between effect size and *SE* and potentially publication bias.

With evidence of an asymmetric distribution, the trim and fill method was used to characterize the impact of observed asymmetry on the overall results (Duval & Tweedie, 2000). This method identifies data points (studies) that contribute to asymmetry and generates a counterpart data point to offset the asymmetry, resulting in a symmetric distribution. The resulting symmetric distribution includes all of the studies from the original corpus combined with hypothetical studies that would be present if the distribution was symmetric. The overall effect size of this combined dataset was compared to the original, overall effect size of the original corpus to estimate the impact potential publication bias had on the original, overall effect size (Rothstein, Sutton, & Borenstein, 2005).

3 | RESULTS

The 99 studies that met the criteria, including effect size calculation and the resulting codes on EBIP type and assessment coverage, are presented in Table S1 in the online supplement. An outlier screening identified one study that is three *SD*s higher from the mean: Tarhan and Acar-Sesen (2013). There are four more studies, Tarhan, Ayyıldız, Ogunc, and Sesen (2013), Acar and Tarhan (2007), Acar and Tarhan, Ayar-Kayali, Urek, and Acar (2008) and Eymur and Geban (2017), that are two *SD*s higher from the average. The analyses that follow have these five studies omitted except where noted. The overall effect size was calculated using a random effects model for each tier of outliers and descriptive statistics of the overall effect sizes are presented in Table 2. The differing approaches to characterizing outliers had no substantive impact on the major conclusions reached.

TABLE 2 Results of outlier screening

Outliers	Number of studies	Weighted mean effect size	Median effect size	*SD*
All studies	99	0.717	0.618	0.818
≤3 *SD* from mean	98	0.685	0.602	0.748
≤2 *SD* from mean	94	0.618	0.568	0.649

3.1 | Effectiveness of EBIPs in chemistry

The overall average effect size of EBIPs in chemistry on students' assessment performance was found to be 0.62. This observed effect size is analogous to a Cohen's d between medium ($d = 0.5$) and large ($d = 0.8$) using Cohen's qualitative descriptors (Cohen, 1988). In short, the research base represented by these 94 studies point to a statistically significant and notably higher chemistry students' tests scores with the use of EBIPs as compared to traditional instruction. The observed effect size falls close to the top end of the range of overall effect sizes from past meta-analyses particular to chemistry: 0.37–0.68 (Apugliese & Lewis, 2017; Leontyev et al., 2017; Warfa, 2016). As noted though, this analysis comprises a broader picture of instructional interventions in chemistry as demonstrated by the relative number of studies. The overall effect size is also slightly greater than the overall effect sizes found in past large-scale meta-analyses in STEM or science education, which range from 0.47 to 0.50 (Freeman et al., 2014; Ruiz-Primo et al., 2011). There is significant variability among the studies with $Q_b = 1{,}174.61$ ($p < .05$), which is expected as studies varied in instructional interventions, assessment types and settings. In considering research methodology, 79 studies used a quasi-experimental methodology comparing established classes or comparison groups of students and 15 studies used an experimental design with random assignment to create classes or comparison groups. Reported effectiveness of pedagogies between methodologies was similar with quasi-experimental average effect size of 0.60 ($SE = 0.05$) versus experimental average effect size of 0.73 ($SE = 0.15$). Given the small sample of experimental studies, research methodology was not considered as a moderator in the ensuing analyses.

3.2 | Relative effectiveness of EBIPs

Studies were demarcated by EBIPs as shown in Table 3. It is evident that there are relatively few studies for each EBIP that meet the criteria for the meta-analysis. The numbers of studies ranged from 7 for PLTL to 15 for Flipped and the *SE* for each of these is substantial, ranging from 0.12 to 0.17. The weighted mean effect size for collaborative learning and PBL studies came up with larger effect sizes than the other EBIPs. The effects size indices for each of these EBIPs exceeds Cohen's description of a large ($d = 0.80$) effect size (1988). Among PLTL, POGIL and Flipped classes in chemistry, the weighted mean effect size indicates that a positive small to medium effect has been realized. It is also worth noting that the weighted mean effect size for POGIL of 0.30 is comparable to the 0.22 results observed in a recent meta-analysis on POGIL implementation across disciplines (Walker & Warfa, 2017). The studies with nonspecified cooperative learning features 33 studies and a larger weighted mean effect size of 0.71.

There is a noticeable variability between and within each EBIP. The confidence intervals for POGIL spans from no effect to medium effect sizes, Flipped from small to medium effect, PLTL

TABLE 3 Impact of moderators (EBIP and assessment coverage) on effect size

	k	Weighted mean effect size	*SE*	95% confidence interval	Q_m (*p*-value, τ^2, I^2)
Types of EBIPs					
Collaborative	13	0.95	0.14	[0.67, 1.23]	102.1 (<.001, 0.189, 91.10%)
PBL	12	0.91	0.15	[0.61, 1.21]	
PLTL	7	0.48	0.17	[0.14, 0.82]	
POGIL	10	0.30	0.15	[0.00, 0.60]	
Flipped	15	0.36	0.12	[0.12, 0.60]	
Nonspecified	33	0.71	0.09	[0.55, 0.89]	
Assessment coverage					
Single topic	49	0.87	0.07	[0.73, 1.01]	174.6 (<.001, 0.160, 89.13%)
Cumulative	24	0.25	0.09	[0.07, 0.43]	
Overall					
Overall	94	0.618	0.05	[0.522, 0.713]	

from small to large effect, and collaborative learning and PBL from approximately medium to large. Each of the EBIPs confidence intervals span positive values substantiating their inclusion as an instructional practice with a demonstrated evidence base of promoting successful student academic performance. The lower bound for the confidence interval of POGIL reaches zero, suggesting that the evidence base is inconsistent, but may be explained by the role of assessment coverage as discussed later. The Q_m statistic observed of 102.1 is statistically significant ($p < .05$) indicating that the type of EBIP explains a portion of the heterogeneity observed among the effect sizes in the corpus. The results in Table 3 indicate that collaborative and PBL instructional practices are expected to offer stronger academic benefits than PLTL, POGIL or Flipped; but such a conclusion is hasty and requires a more in-depth look at the studies.

Studies using single-topic assessment and studies using cumulative assessment topics are each well represented within the corpus of studies as shown in Table 3. Studies with single-topic assessments have a weighted mean effect size of 0.87 in contrast to studies using cumulative topic assessments averaging 0.25. The confidence intervals of single-topic and cumulative do not overlap, indicating that EBIPs have a demonstrably larger impact on student performance when measured by narrowly defined assessments spanning a small number of topics than on cumulative assessments spanning an entire term, in line with findings from an earlier meta-analysis (Apugliese & Lewis, 2017).

Given the role of assessment coverage in impacting observed effect sizes, the data for each type of EBIPs was demarcated based on assessment coverage in Table 4. Of the 99 studies, 12 studies used both single-topic and cumulative assessments to evaluate the intervention (referred to as split studies), 8 studies reported a total score that combined both types of assessments and 1 study did not include sufficient information to code assessment type. These 21 studies for each category were not considered in Table 4 but an analysis that includes the split studies is presented in the online supplement with no substantive change in interpretation. The demarcation by assessment coverage explains some of the trends observed among the EBIPs. First, the higher overall average of collaborative learning and PBL is partially explained since a large majority of the studies for those two EBIPs (11 out of 13 for collaborative and 9 out of

TABLE 4 Interaction of EBIP and assessment coverage type

	Overall			Single topic			Cumulative		
Type of EBIPs	***k***	**Mean**	***SE***	***k***	**Mean**	***SE***	***k***	**Mean**	***SE***
Collaborative	13	0.95	0.14	11	1.05	0.24	1	0.61	0.51
PBL	12	0.91	0.15	9	1.24	0.27	2	−0.19	0.30
PLTL	7	0.48	0.17		N/A		4	0.14	0.16
POGIL	10	0.30	0.15	2	0.87	0.51	6	0.15	0.15
Flipped	15	0.36	0.12	5	0.48	0.35	3	0.31	0.20
Nonspecified	33	0.71	0.09	22	0.78	0.12	6	0.44	0.14
Overall	94	0.62	0.05	49	0.90	0.10	24	0.24	0.06

TABLE 5 Descriptive statistics for setting sizes by EBIP

Type of EBIPs	Median setting size	Range	Effect size
Collaborative	32	16–81	0.95
PBL	35.5	20–79	0.91
Nonspecified	53	17–3,174	0.71
Flipped	66	7–864	0.36
POGIL	109.5	26–193	0.30
PLTL	353	35–1,037	0.48

12 for PBL) used single-topic assessments. PBL had the highest weighted mean effect size ($g = 1.24$) among single-topic assessments, and other EBIPs such as PLTL (no studies) and POGIL (two studies) have too few studies to make a comparison. While the overall effect for PBL and collaborative appears to be inflated owing to single-topic assessments, PLTL and POGIL may be weighted down by their high rate of cumulative assessments. The importance of this moderator is demonstrated with the large swings evident in PBL and POGIL across assessment type. Among the EBIPs, studies on flipped teaching appear relatively stable across assessment coverage but even then span from small to medium impact. Ultimately, the relative effectiveness of collaborative and PBL in comparison to other EBIPs is tempered by the distribution of assessment coverage used and the number of studies prevents definitive comparisons of EBIPs while controlling for assessment coverage.

Variation across different EBIPs is also partially explained by the setting size of the study. Setting size serves as a proxy for class size as studies with larger setting sizes tend to study larger class sizes; some studies did not report class size preventing recording actual class size across all studies. Descriptive statistics on the sample size of the treatment group for each EBIP is presented in Table 5 along with the overall weighted mean effect size. There is an inverse relationship observed between setting size and effect size. This matches the previous finding by Freeman et al. (2014) and Warfa (2016) that alternative pedagogies have a larger impact when class size is small. The median setting size for collaborative and PBL is quite smaller than the rest of the EBIPs particularly Flipped, POGIL and PLTL indicating that setting size serves as an additional confounding variable in comparing EBIPs. In summary, the relative effectiveness of

each EBIP cannot be definitively determined with this corpus of data owing to the potential confounding effects of assessment coverage and class size.

Additional moderators likely also play a role in understanding the evidence-base including whether the pedagogy was implemented in a chemistry laboratory course versus a conventional classroom or in a postsecondary versus secondary institution. The strong majority, 84 of the 94 studies, were conducted in a conventional classroom with an average effect size of 0.58 ($SE = 0.05$). Studies conducted in a chemistry laboratory course were far less common, including 10 studies with an average effect size of 0.92 ($SE = 0.19$). Studies were more evenly split between postsecondary versus secondary institutions though postsecondary studies, including professional schools, comprise the majority of the corpus. Of the 94 studies, 62 took place at a postsecondary institution with an average effect size of 0.50 ($SE = 0.05$). In contrast, 32 studies at a secondary school had an average effect size of 0.87 with ($SE = 0.10$), with 27 of these 32 studies using single-topic assessments. As before, the size of the corpus prevents exploring the relative effectiveness of EBIPs within each of these research settings.

3.3 | Investigation of publication bias

For the purpose of determining whether publication bias was present among the corpus of studies, a funnel plot was created using comprehensive meta-analysis version 3.0 (Borenstein, Hedges, & Higgins, 2013). The funnel plot is shown in Figure 2, with each circle representing a study, and was visually inspected for symmetry. Asymmetry, indicative of publication bias, is visibly evident in the funnel plot with studies on the right side of the plot disproportionately appearing toward the bottom of the plot. This trend matches the aforementioned finding that smaller sample sizes (larger *SE* on the funnel plot in Figure 2) tended to have larger effect sizes. Follow-up tests both supported an interpretation of asymmetry matching the visual inspection; rank correlation test (Kendal tau = 0.32, $p < .05$) and Egger's regression test (intercept = 3.22, $p < .05$) each resulted in a statistically significant coefficient rejecting the null hypothesis of a symmetric distribution.

The trim and fill method was used to assess the impact of asymmetry on the weighted average of the effect size of this corpus of studies with results shown in Figure 3. The trim and fill method is intended to simulate a symmetric distribution and then describe the weighted average effect size of the hypothetical symmetric distribution. If the adjusted effect size is similar to

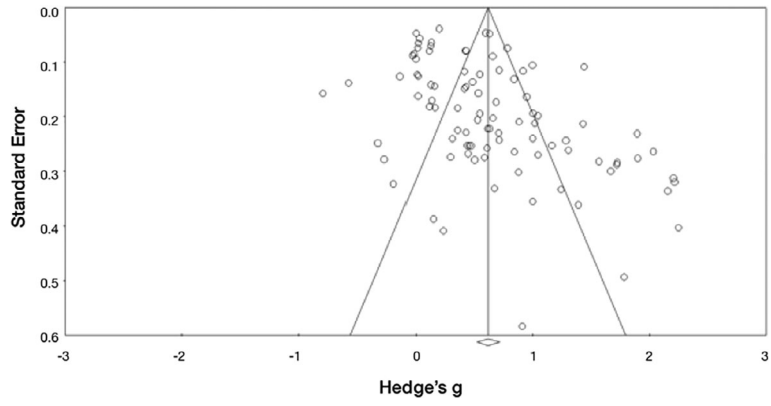

FIGURE 2 Funnel plot shows an asymmetric distribution

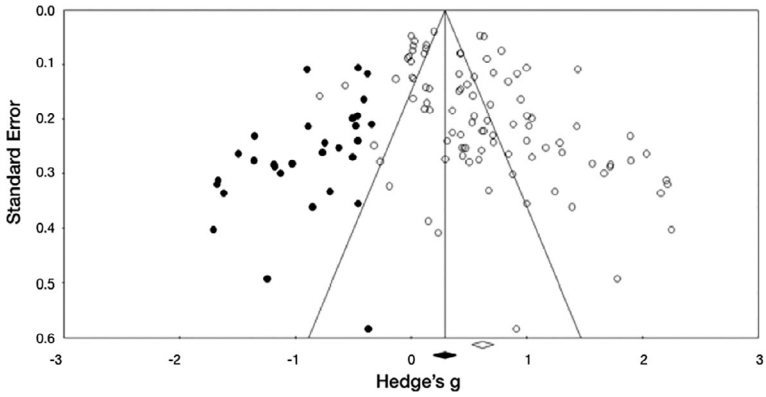

FIGURE 3 Funnel plot with trim and fill

the original effect size the effect of publication bias can be described as negligible; if the adjusted effect size is notably different from the original effect size yet the interpretation of both effect sizes would remain consistent the effect of publication bias is moderate; and if the adjusted effect size would change the conclusions reached the publication bias can be described as severe (Kepes, Banks, McDaniel, & Whetzel, 2012). The trim and fill method indicated a noteworthy shift downward, as the asymmetry is indicative of an inflated value of the weighted average effect size. The calculated weighted average effect size through the trim and fill method was 0.29 (95% confidence interval of 0.19–0.39), which can be described as a moderate decrease from the original value of 0.62 (95% confidence interval of 0.52–0.71). The adjusted value of 0.29 should be interpreted with caution. The overall corpus of studies was found to be heterogeneous with $Q_b = 1{,}174.16$ ($p < .05$). With high heterogeneity in a dataset, the trim and fill method likely underestimates the overall treatment effect (Peters, Sutton, Jones, Abrams, & Rushton, 2007; Terrin, Schmid, Lau, & Olkin, 2003). One interpretation of the asymmetric distribution is that studies with small sample sizes and small effect sizes, a combination that would fail to generate statistical significance, were less likely to be submitted or accepted for publication. An alternative hypothesis is that the asymmetry demonstrates an authentic relationship in the data where effectiveness of EBIPs diminishes with larger class sizes owing to logistical or instructional challenges. For example, enacting an EBIP with a large class size may limit the extent or quality of individualized student feedback, which may be necessary for academic gains.

To explore this hypothesis, a content review of the eight studies reporting setting sizes greater than 400 was conducted. Five studies (Eichler & Peeples, 2016; He, 2016; Lewis, 2011; Robert et al., 2016; Tien, Roth, & Kampmeier, 2002) used peer leaders or teaching assistants to facilitate interactions within large classes. Of the remaining three studies, two reported class sizes of approximately 100 students (Baepler et al., 2014; Casadonte, 2016) and the remaining study (Talanquer & Pollard, 2017) a class size of 250 students. Two (He, 2016; Talanquer & Pollard, 2017) of the eight studies make explicit mention of the challenges in implementation with a large class describing difficulties in ensuring student preparation, promoting student engagement and providing feedback on misconceptions. It is also noted that He (2016) was the only study of the five with teaching assistants to not mention the number of assistants present. Thus, it may be that the use of peer leaders or teaching assistants with a smaller student to assistant ratio may mitigate the challenges of large classes, but a large student to assistant ratio or the absence of assistants poses substantive challenges in implementation. It is also possible that

TABLE 6 Publication bias results by outlier decision

Outliers screening procedure	*k*	Rank correlation test	Egger's regression test	Weighted effect size (after trim and fill; before)
Entire corpus	99	Kendall's tau = 0.320 $p < .05$	Intercept = 3.220 $p < .05$	(0.310; 0.717)
≤3 *SD* from mean	98	Kendall's tau = 0.306 $p < .05$	Intercept = 3.032 $p < .05$	(0.300; 0.675)
≤2 *SD* from mean	94	Kendall's tau = 0.259 $p < .05$	Intercept = 2.650 $p < .05$	(0.292; 0.618)

both publication bias and challenges with implementation in large classes combine to create the asymmetry observed.

Thus, a suggested interpretation for the average impact of EBIPs in chemistry while taking into account possible publication bias is that the actual average would lie within the range of 0.29–0.62, with the lower bound from the trim and fill approach and the upper bound unadjusted from the original weighted average. The entirety of this range is positive and exceeds a small effect size indicating that the evidence base of EBIPs promoting student success is maintained. In summary, the evidence base for EBIPs is likely overstated owing to publication bias but the evidence base remains robust enough to warrant adoption.

The decision for outlier screening was revisited to determine the impact this decision had on publication bias. For each outlier removal procedure, the funnel plot was developed and the subsequent tests (rank correlation test, Egger's regression test, trim and fill method) were conducted. The result indicated a similar pattern where the effect size decreased to 0.31 and 0.30 for the entire corpus and three *SD*s from the mean respectively as demonstrated in Table 6. The outlier decision appears to have minimal impact on the publication bias analysis and would not alter the interpretation of the results.

4 | DISCUSSION

The overall effectiveness and the effectiveness demarcated by EBIP strategy indicate consistent learning gains in enacting EBIPs within chemistry instruction thereby supporting the adoption of any of the EBIPs described herein. One of the original goals of the meta-analysis was to conduct a comparison of the relative effectiveness of each EBIP. Such a comparison has been explicitly called for in recent reviews of science education research (Freeman et al., 2014; National Research Council, 2012 p. 137). The comparison of relative effectiveness for each EBIP was hindered by confounding variables in the form of cumulative versus single-topic assessments and setting size and there were insufficient studies to control for these confounding variables. Even so, the analysis offers insight into the current evidence-base and limitations therein for each EBIP, which can inform instructional decisions to adopt and directions for future research.

The generic EBIP of nonspecified cooperative learning features the most substantive evidence-base with medium to large effect sizes across single-topic and cumulative assessment types and across a range of setting sizes. An instructional decision to enact cooperative learning is therefore supported across a variety of instructional settings. Collaborative and PBL feature

the strongest effect sizes among the EBIPs evaluated but the research base is limited to primarily single-topic assessments and smaller setting sizes. Among the EBIPs originating within chemistry, POGIL has primarily been evaluated with cumulative assessments and smaller setting sizes showing moderately higher student academic performance than control groups. Thus, instruction with smaller class sizes appears likely to result in sizable observed benefits from collaborative, PBL and POGIL but the evidence-base does not yet warrant implementation in large classes. Among the three, POGIL may have the most promising case for moving to large classes as four studies had setting sizes greater than 150 and effect sizes ranging from 0.00 to 0.71. In large classes, PLTL has the strongest evidence base with five of seven studies reporting setting sizes greater than 200 and effect sizes ranging from 0.02 to 0.84. This matches the scalability of PLTL where larger class sizes can be accommodated by increasing the number of peer leaders supporting implementation (Robert et al., 2016). Future research on the effectiveness of POGIL, PBL and collaborative learning in large classes and PLTL in small classes is still needed and could also include qualitative investigations into how class size influences the implementation of these approaches.

Flipped learning has an emergent research base with 15 studies reported all since 2013 and 11 of the studies published in 2016 or 2017. The studies span single-topic and cumulative assessment types and a range of research settings with a median setting size of 66 and five studies of setting sizes with more than 300 students. The evidence-base for flipped learning mirrors that of nonspecified cooperative learning although with approximately half the studies included and an overall effect size considerably lower than nonspecified cooperative learning (0.36 vs. 0.71). The difference may be the result of the variation in flipped learning as it provides less direction into how to enact in-class active learning once instruction has been moved out of class (see literature review in Robert et al., 2016).

Overall, EBIPs have shown less effectiveness when measured with a cumulative exam relative to single-topic exam and future research exploring why this difference arises would be informative. One potential explanation for this difference is that EBIPs primarily promote short-term understanding but are less effective at promoting long-term understanding. Another explanation is that cumulative assessments are more likely than single topic assessments to include some items that were not presented via EBIP. Single topic assessments by definition are more focused by topic than cumulative assessments. For example, some studies used an EBIP to target a particular topic and evaluated the effectiveness with a concept inventory on the same topic (Acar & Tarhan, 2008; Doymus, 2007; Doymus, 2016). In contrast, studies using an EBIP throughout a semester and evaluated the effectiveness with a cumulative assessment may employ EBIP with a majority of topics but employ traditional instruction with a subset of select topics. In the evaluation, assessment items related to these select topics within a cumulative exam would be expected to show little or no difference between pedagogies and lower the overall observed effect size. Better understanding of the underlying reasons for the differences between single topic and cumulative assessments is necessary to promote the robustness of EBIPs' evidence-base across assessment types.

The analysis of publication bias within the corpus of articles shows that the overall effect calculated by meta-analysis may be overstated. The trend observed in the data was a disproportionate incidence of larger effect sizes observed among studies with smaller sample sizes and smaller effect sizes observed among studies with larger sample sizes. This trend raises the possibility that a group of studies with smaller effect size and smaller sample size, a combination that would tend toward a failure to show statistical significance, were conducted but not published. Researchers in the field may be less likely to attempt to publish these findings or reviewers and editors in the field may be less likely to accept these findings for publication. The importance of publishing null results to reach an accurate measure of the impact of alternative pedagogies

needs to be emphasized. An alternative explanation is that there may be a relationship between class size and the effectiveness of the pedagogies investigated. This explanation furthers the aforementioned need to investigate the role of class sizes on EBIP implementation. In spite of the publication bias evidence, the findings remain supportive of the use of EBIPs in chemistry teaching associated with demonstrably higher academic performance. Additionally, it has been argued that meta-analyses can provide a benchmark for evaluating future work in the field (Lipsey et al., 2012). The range of 0.29–0.62 can therefore serve as a minimum and maximum expected effectiveness of EBIPs in chemistry instruction and can serve to gauge the relative effectiveness of future implementations of alternative instructional practices.

To provide greater context to the results reported, we sought to better understand the instruction within the control group, which serves as the comparison condition, for studies within the corpus. A content analysis was performed on the 96 studies that have a unique control group. As mentioned, three studies had two unique experimental conditions and each contributed two effect sizes to the analysis but had only one control group. Additionally, 10 studies took place in a laboratory course setting with a control group of a laboratory course, 8 of these 10 studies described the comparison lab course as traditional. Within the 86 studies taking place in a classroom eight studies offered no description of the instruction taking place in the control group. Analyzing the remaining 78 studies, 58 studies explicitly described relying on lecture or didactic instruction, the most common description of the control group. Nearly as frequently, 57 studies describe instruction as traditional or conventional, implying a continuation of past practices. Combined, 68 of the 78 studies were described as using traditional instruction, lecture or used both traditional and lecture to describe the control. Thirty-two of the 78 studies described students working on problems individually or having assigned homework, but most of these (25 of the 32) also reported lecture instruction. Similarly, 22 studies described teachers modeling problem solving, asking or answering student questions or including a recitation session and 19 of these 22 also relied on lecture instruction. Studies also described supplementing lecture instruction in the control group with clicker use (seven studies), demonstrations (six) and group work (four). The primary control group condition that did not mention lecture instruction was the use of computer based instruction (four studies). Distinctively, one study used project-based learning as a control group (Paristiowati, Erdawati, & Nurtanti, 2017) to compare with project-based learning via the flipped model; another study used guided-inquiry as a control group (Paristiowati, Fitriani, & Aldi, 2017) to compare with to inquiry via the flipped model. In summary, the strong majority of studies relied on lecture-based instruction in the control group, with some variety in how lecture-based instruction was supplemented. While it is not possible within the corpus to characterize the exact extent lecturing was taking place in each control group it is clear that this corpus of studies describes moving away from lecture instruction and has resulted in a demonstrable, positive effect on student academic performance.

Limitations for this meta-analysis include the potential for additional confounding variables present among characteristics that were not coded. In particular fidelity of implementation, the extent an instructor enacted the critical criteria described by the EBIP designer, was not measureable by review of the literature. In other words, while there undoubtedly exists variation in the enactment of each EBIP across the set of studies, there was no reliable way to demarcate this variation without additional data sources including instructor interviews or on-site observations. Additionally, this study could not examine all evidence-based instructional practices in chemistry and the use of meta-analytic methodology limited the evidence-base to that generated through studies using experimental or quasi-experimental comparisons on students' academic performance. Finally, the techniques used to investigate publication bias are not

sensitive to p-hacking, the use of multiple analyses on a dataset to eventually arrive at statistical significance. Investigating the presence of p-hacking, see Simonsohn, Nelson, and Simmons (2014) for more information, are warranted in future work.

5 | CONCLUSIONS

This study sought to provide a discipline specific synthesis of EBIPs through meta-analysis and to that end the identified literature comprised the broadest view of experimental and quasi-experimental, chemistry-specific studies to date. The results showed that classes using the reviewed EBIPs have demonstrably higher scores on chemistry student academic performance. Assessment topic coverage and setting size within the studies emerged as relevant moderators of impact and prevented making definitive conclusions of the relative impact of each EBIP. The distribution of studies in terms of setting size to effect size was asymmetrical providing the possibility that either studies with small sample size and small effect size were not published (publication bias) or that large class sizes feature unique challenges that hinder EBIP effectiveness. Modeling hypothesized studies to generate a symmetric distribution provides a range for the overall weighted effect size of 0.29–0.62 indicative that the evidence base for EBIPs is robust and warrants adoption.

ACKNOWLEDGMENTS

Partial support for this work was provided by the National Science Foundation's Improving Undergraduate STEM Education (IUSE) program under DUE-1712164. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors also acknowledge Michael Brannick for reviewing an earlier draft of this manuscript and Andrew Apugliese for assistance in screening the literature.

ORCID

Scott E. Lewis https://orcid.org/0000-0002-6899-9450

REFERENCES

*Asterisk indicates articles that are part of the corpus of studies included.*Acar, B., & Tarhan, L. (2007). Effect of cooperative learning strategies on students' understanding of concepts in electrochemistry. *International Journal of Science and Mathematics Education*, *5*, 349–373. https://doi.org/10.1007/s10763-006-9046-7

*Acar, B., & Tarhan, L. (2008). Effects of cooperative learning on students' understanding of metallic bonding. *Research in Science Education*, *38*, 401–420. https://doi.org/10.1007/s11165-007-9054-9

*Akinyele, A. F. (2010). Peer-led team learning and improved performance in an allied health chemistry course. *The Chemical Educator*, *15*, 8.

*Aldridge, J. N. (2011). From access to success in science: An academic-student affairs intervention for undergraduate freshmen biology students. *Dissertation Abstracts International*, *73*(02) Retrieved from ProQuest Dissertations & Theses database. (AAT 3478721).

*Allen, D. A. (2003). *Development and assessment of an active learning environment: cAcL2 concept advancement through chemistry laboratory-lecture*. (Doctoral dissertation). North Carolina State University. Retrieved from ProQuest Dissertations & Theses database. (3098927)

Apugliese, A., & Lewis, S. E. (2017). Impact of instructional decisions on the effectiveness of cooperative learning in chemistry through meta-analysis. *Chemistry Education Research and Practice*, *18*, 271–278. https://doi.org/10.1039/C6RP00195E

*Awan, R. U. N., Hussain, H., & Anwar, N. (2017). Effects of problem based learning on students' critical thinking skills, attitudes towards learning and achievement. *Journal of Educational Research*, *20*, 28–41.

*Baepler, P., Walker, J. D., & Driessen, M. (2014). It's not about seat time: Blending, flipping, and efficiency in active learning classrooms. *Computers & Education*, *78*, 227–236. https://doi.org/10.1016/j.compedu.2014.06.006

*Baran, M. (2016). Teaching gases through problem-based learning. *Journal of Education and Training Studies*, *4* 281–294. https://doi.org/10.11114/jets.v4i4.1498

Barkley, E. F., Cross, K. P., & Major, C. H. (2014). *Collaborative learning techniques: A handbook for college faculty*. San Francisco, CA: Jossey-Bass.

*Barthlow, M. J., & Watson, S. B. (2014). The effectiveness of process-oriented guided inquiry learning to reduce alternative conceptions in secondary chemistry: Effectiveness of POGIL on alternative conceptions. *School Science and Mathematics*, *114*, 246–255. https://doi.org/10.1111/ssm.12076

Becker, B. J., Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Sussex, England: John Wiley & Sons, Ltd.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*, 1088–1101. https://doi.org/10.2307/2533446

*Bernard, P., Broś, P., & Migdał-Mikuli, A. (2017). Influence of blended learning on outcomes of students attending a general chemistry course: Summary of a five-year-long study. *Chemistry Education Research and Practice*, *18*, 682–690. https://doi.org/10.1039/c7rp00040e

*Bilgin, I. (2006). Promoting pre-service elementary students' understanding of chemical equilibrium through discussions in small groups. *International Journal of Science and Mathematics Education*, *4*, 467–484. https://doi.org/10.1007/s10763-005-9015-6

*Bilgin, I. (2009). The effects of guided inquiry instruction incorporating a cooperative learning approach on university students' achievement of acid and bases concepts and attitude toward guided inquiry instruction. *Scientific Research and Essays*, *4*, 1038–1046.

*Bilgin, İ., & Geban, Ö. (2006). The effect of cooperative learning approach based on conceptual change condition on students' understanding of chemical equilibrium concepts. *Journal of Science Education and Technology*, *15*, 31–46. https://doi.org/10.1007/s10956-006-0354-z

*Bilgin, I., Šenocalk, E., & Sözbilir, M. (2009). The effects of problem-based learning instruction on university students' performance of conceptual and quantitative problems in gas concepts. *Eurasia Journal of Mathematics, Science & Technology Education*, *5*, 153–164.

Borenstein, M., Hedges, L. V., & Higgins, J. P. T. (2013). In H. R. Rothstein (Ed.), *Comprehensive meta-analysis version 3 [Software]*. Englewood, NJ: Biostat.

*Bramaje, G. P., & Espinosa, A. A. (2013). Peer-led team learning approach: Effects on students' conceptual understanding and attitude towards chemistry. *International Journal of Teaching and Learning*, *5*, 55–77.

*Brown, S. D. (2010). A process-oriented guided inquiry approach to teaching medicinal chemistry. *American Journal of Pharmaceutical Education*, *74*, 121. https://doi.org/10.5688/aj7407121

*Cam, A., & Geban, Ö. (2013). Effectiveness of case-based learning instruction on dtudents' understanding of solubility equilibrium concepts. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *44*, 97–108.

*Canelas, D. A., Hill, J. L., & Novicki, A. (2017). Cooperative learning in organic chemistry increases student assessment of learning gains in key transferable skills. *Chemistry Education Research and Practice*, *18*, 441–456. https://doi.org/10.1039/c7rp00014f

*Casadonte, D. (2016). *The effectiveness of course flipping in general chemistry—Does it work? ACS symposium series, 1228, 19–37*. Washington, DC: Oxford University Press.

*Chase, A., Pakhira, D., & Stains, M. (2013). Implementing process-oriented, guided-inquiry learning for the first time: Adaptations and short-term impacts on students' attitude and performance. *Journal of Chemical Education*, *90*, 409–416. https://doi.org/10.1021/ed300181t

*Chen, Y. C. (2013). *Learning protein structure with peers in an AR-enhanced learning environment*. University of Washington: (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses database. (3588651)

*Christiansen, M. A. (2014). Inverted teaching: Applying a new pedagogy to a university organic chemistry class. *Journal of Chemical Education*, *91*, 1845–1850. https://doi.org/10.1021/ed400530z

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

*Crimmins, M. T., & Midkiff, B. (2017). High structure active learning pedagogy for the teaching of organic chemistry: Assessing the impact on academic outcomes. *Journal of Chemical Education*, *94*, 429–438. https://doi.org/10.1021/acs.jchemed.6b00663

*Díaz-Vázquez, L. M., Montes, B., Echevarría Vargas, I., Hernandez-Cancel, G., Gonzalez, F., Molina, A., ... Griebenow, K. (2012). An investigative, cooperative learning approach for general chemistry laboratories. *International Journal for the Scholarship of Teaching and Learning*, *6*, 1–14. https://doi.org/10.20429/ijsotl.2012.060220

*Ding, N., & Harskamp, E. G. (2011). Collaboration and peer tutoring in chemistry laboratory education. *International Journal of Science Education*, *33*, 839–863. https://doi.org/10.1080/09500693.2010.498842

*Doymus, K. (2007). Effects of a cooperative learning strategy on teaching and learning phases of matter and one-component phase diagrams. *Journal of Chemical Education*, *84*, 1857. https://doi.org/10.1021/ed084p1857

*Doymus, K. (2008a). Teaching chemical bonding through jigsaw cooperative learning. *Research in Science & Technological Education*, *26*, 47–57. https://doi.org/10.1080/02635140701847470

*Doymus, K. (2008b). Teaching chemical equilibrium with the jigsaw technique. *Research in Science Education*, *38*, 249–260. https://doi.org/10.1007/s11165-007-9047-8

*Doymus, K., Karacop, A., & Simsek, U. (2010). Effects of jigsaw and animation techniques on students' understanding of concepts and subjects in electrochemistry. *Educational Technology Research and Development*, *58*, 671–691. https://doi.org/10.1007/s11423-010-9157-2

*Doymus, K., Simsek, U., & Karacop, A. (2009). The effects of computer animations and cooperative learning methods in micro, macro and symbolic level learning of states of matter. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, *36*, 109–128.

Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*, 89–98. https://doi.org/10.1080/01621459.2000.10473905

Eberlein, T., Kampmeier, J., Minderhout, V., Moog, R. S., Platt, T., Varma-Nelson, P., & White, H. B. (2008). Pedagogies of engagement in science: A comparison of PBL, POGIL, and PLTL. *Biochemistry and Molecular Biology Education*, *36*, 262–273. https://doi.org/10.1002/bmb.20204

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *The BMJ*, *315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629

*Eichler, J. F., & Peeples, J. (2016). Flipped classroom modules for large enrollment general chemistry courses: A low barrier approach to increase active learning and improve student grades. *Chemistry Education Research and Practice*, *17*, 197–208. https://doi.org/10.1039/c5rp00159e

*Eymur, G., & Geban, Ö. (2017). The collaboration of cooperative learning and conceptual change: Enhancing the students' understanding of chemical bonding concepts. *International Journal of Science and Mathematics Education*, *15*, 853–871. https://doi.org/10.1007/s10763-016-9716-z

*Fakomogbon, M. A., & Bolaji, H. O. (2017). Effects of collaborative learning styles on performance of students in a ubiquitous collaborative mobile learning environment. *Contemporary Educational Technology*, *8*, 268–279.

Farrell, J. J., Moog, R. S., & Spencer, J. N. (1999). A guided-inquiry general chemistry course. *Journal of Chemical Education*, *76*, 570. https://doi.org/10.1021/ed076p570

*Foley, K., & O'Donnell, A. (2002). Cooperative learning and visual organisers: Effects on solving mole problems in high school chemistry. *Asia Pacific Journal of Education*, *22*, 38–50. https://doi.org/10.1080/0218879020220105

*Frailich, M., Kesner, M., & Hofstein, A. (2009). Enhancing students' understanding of the concept of chemical bonding by using activities provided on an interactive website. *Journal of Research in Science Teaching*, *46*, 289–310. https://doi.org/10.1002/tea.20278

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*, 8410–8415. https://doi.org/10.1073/pnas.1319030111

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of Educational Research*, *75*, 27–61. https://doi.org/10.3102/00346543075001027

*Glynn, J., Jr. (2013). *The effects of a flipped classroom on achievement and student attitudes in secondary chemistry*. (MS dissertation). Montana: Montana State University Bozeman. Retrieved from ProQuest Dissertations & Theses database. (AAT 3478721)

*Goeden, T. J., Kurtz, M. J., Quitadamo, I. J., & Thomas, C. (2015). Community-based inquiry in allied health biochemistry promotes equity by improving critical thinking for women and showing promise for increasing content gains for ethnic minority students. *Journal of Chemical Education*, *92*, 788–796. https://doi.org/10.1021/ed400893f

*Günter, T., Akkuzu, N., & Alpat, Ş. (2017). Understanding 'green chemistry' and 'sustainability': An example of problem-based learning (PBL). *Research in Science & Technological Education*, *35*, 500–520. https://doi.org/10.1080/02635143.2017.1353964

*Hagen, J. P. (2000). Cooperative learning in organic II. Increased retention on a commuter campus. *Journal of Chemical Education*, *77*, 1441. https://doi.org/10.1021/ed077p1441

*He, W. (2016). *Investigation of the effects of flipped instruction on student exam performance, motivation and perceptions* (Doctoral dissertation). UC Irvine. Retrieved from ProQuest Dissertations & Theses database. (10168597).

*Hein, S. M. (2012). Positive impacts using POGIL in organic chemistry. *Journal of Chemical Education*, *89*, 860–864. https://doi.org/10.1021/ed100217v

*Hemraj-Benny, T., & Beckford, I. (2014). Cooperative and inquiry-based learning utilizing art-related topics: Teaching chemistry to community college nonscience majors. *Journal of Chemical Education*, *91*, 1618–1622. https://doi.org/10.1021/ed400533r

*Ibraheem, T. L. (2011). Effects of two modes of student teams-achievement division strategies on senior secondary school students' learning outcomes in chemical kinetics. *Asia-Pacific Forum on Science Learning & Teaching*, *12*, 1–21.

*Jiang, B. (2014). Web-based cooperative learning in college chemistry teaching. *International Journal of Emerging Technologies in Learning*, *9*, 45. https://doi.org/10.3991/ijet.v9i2.3224

*Joel, G. E., Kamji, D. T., & Godiya, E. E. (2016). Enhancing pre-degree chemistry students' conceptual understanding of rates of chemical reactions through cooperative learning strategy. *International Journal of Innovative Research and Development*, *5*, 322–327.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1998). Cooperative learning returns to college. What evidence is there that it works? *Change: The Magazine of Higher Learning*, *30*, 26–35. https://doi.org/10.1080/00091389809602629

*Jong, J. P. (2016). The effect of a blended collaborative learning environment in a small private online course (SPOC): A comparison with a lecture course. *Journal of Baltic Science Education*, *15*, 194–203.

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods*, *15*, 624–662.

*Khan, G. N., & Inamullah, H. M. (2011). Effect of Student's Team Achievement Division (STAD) on academic achievement of students. *Asian Social Science*, *7*, 211. https://doi.org/10.5539/ass.v7n12p211

*Kırık, Ö. T., & Boz, Y. (2012). Cooperative learning instruction for conceptual change in the concepts of chemical kinetics. *Chemistry Education Research and Practice*, *13*, 221–236. https://doi.org/10.1039/c1rp90072b

*Kiste, A. L., Scott, G. E., Bukenberger, J., Markmann, M., & Moore, J. (2017). An examination of student outcomes in studio chemistry. *Chemistry Education Research and Practice*, *18*, 233–249. https://doi.org/10.1039/C6RP00202A

*Koc, Y., Doymuş, K., Karaçöp, A., & Şimşek, Ü. (2010). The effects of two cooperative learning strategies on the teaching and learning of the topics of chemical kinetics. *Journal of Turkish Science Education*, *7*, 52–65.

Leontyev, A., Chase, A., Pulos, S., & Varma-Nelson, P. (2017). Assessment of the effectiveness of instructional interventions using a comprehensive meta-analysis package. In T. Gupta (Ed.), *ACS symposium series* (Vol. 1260, pp. 117–132). Washington, DC: Oxford University Press. https://doi.org/10.1021/bk-2017-1260.ch008

*Lewis, S. E. (2011). Retention and reform: An evaluation of peer-led team learning. *Journal of Chemical Education*, *88*, 703. https://doi.org/10.1021/ed100689m

*Lewis, S. E., & Lewis, J. E. (2005). Departing from lectures: An evaluation of a peer-led guided inquiry alternative. *Journal of Chemical Education*, *82*, 135–139. https://doi.org/10.1021/ed082p135

*Lewis, S. E., & Lewis, J. E. (2008). Seeking effectiveness and equity in a large college chemistry course: An HLM investigation of peer-led guided inquiry. *Journal of Research in Science Teaching*, *45*, 794–811. https://doi.org/10.1002/tea.20254

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Frey, K., Cole, M. W., … Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms, (NCSER 2013–3000)*. Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

*Lyon, D. C. (2002). *Achievement through small-group discussion sessions in large general chemistry lecture classes with the aid of undergraduate peer teaching assistants*. (PhD Dissertation). Austin, TX: The University of Texas. Retrieved from ProQuest Dissertations & Theses database. (3099487)

Mack, M. R., Hensen, C., & Barbera, J. (2019). Metrics and methods used to compare student performance data in chemistry education research articles. *Journal of Chemical Education*, *96*, 401–413. https://doi.org/10.1021/acs.jchemed.8b00713

Minderhout, V., & Loertscher, J. (2007). Lecture-free biochemistry: A process oriented guided inquiry approach. *Biochemistry and Molecular Biology Education*, *35*, 172–180. https://doi.org/10.1002/bmb.39

*Mitchell, Y. D., Ippolito, J., & Lewis, S. E. (2012). Evaluating peer-led team learning across the two semester general chemistry sequence. *Chemistry Education Research and Practice*, *13*, 378–383. https://doi.org/10.1039/C2RP20028G

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moog, R. S., & Spencer, J. N. (2008). POGIL: An overview. In R. S. Moog & J. N. Spencer (Eds.), *Process oriented guided inquiry learning (POGIL). ACS symposium series* (Vol. 994, pp. 1–13). Washington, DC: Oxford University Press. https://doi.org/10.1021/bk-2008-0994.ch001

*Murphy, K. L., Picione, J., & Holme, T. A. (2010). Data-driven implementation and adaptation of new teaching methodologies. *Journal of College Science Teaching*, *40*, 80–86.

National Research Council. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. Washington, DC: National Academies Press. https://doi.org/10.17226/13362

*Ochonogor, C. (2011). Beyond the usual approach of chemistry teaching in high schools. *US-China Education Review B*, *5*, 643–653.

*Ojennus, D. D. (2016). Assessment of learning gains in a flipped biochemistry classroom: Assessment of learning gains. *Biochemistry and Molecular Biology Education*, *44*, 20–27. https://doi.org/10.1002/bmb.20926

*Olakanmi, E. E. (2017). The effects of a flipped classroom model of instruction on students' performance and attitudes towards chemistry. *Journal of Science Education and Technology*, *26*, 127–137. https://doi.org/10.1007/s10956-016-9657-x

*Own, Z. Y., Chen, D. U., & Chiang, H. R. (2010). A study on the effect of using problem-based learning in organic chemistry for web-based learning. *International Journal of Instructional Media*, *37*, 417–431.

*Özden, M. (2009). Enhancing prospective teachers' development through problem-based learning in chemistry education. *Asian Journal of Chemistry*, *21*, 13.

*Paristiowati, M., Erdawati, & Nurtanti, A. (2017). The effect of flipped classroom-project based learning model and learning independence toward students' achievement in chemical bonding: Case study in SMA Santa Ursula Jakarta. In *Proceedings of the 2017 international conference on education and E-learning—ICEEL 2017* (pp. 22–25). Bangkok, Thailand: ACM Press. https://doi.org/10.1145/3160908.3160915

*Paristiowati, M., Fitriani, E., & Aldi, N. H. (2017). The effect of inquiry-flipped classroom model toward students' achievement on chemical reaction rate (p. 030006). *Presented at the 4th International Conference on Research, Implementation, and Education of Mathematics and Science (4TH ICRIEMS): Research and Education for Developing Scientific Attitude in Sciences and Mathematics, Yogyakarta, Indonesia*. doi:https://doi.org/10.1063/1.4995105

*Partanen, L. (2016). Student oriented approaches in the teaching of thermodynamics at universities—Developing an effective course structure. *Chemistry Education Research and Practice*, *17*, 766–787. https://doi.org/10.1039/c6rp00049e

*Perry, M. D., & Wight, R. D. (2008). Using an ACS general chemistry exam to compare traditional and POGIL instruction. In R. S. Moog & J. N. Spencer (Eds.), *Process oriented guided inquiry learning (POGIL). ACS symposium series 994* (pp. 240–247). Washington, DC: Oxford University Press.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, *26*, 4544–4562. https://doi.org/10.1002/sim.2889

*Poon, T., & Rivera, J. (2015). The flipped classroom as an approach for improving student learning and enhancing instructor experiences in organic chemistry. In K. Daus & R. Rigsby (Eds.), *The promise of chemical education: Addressing our students' needs. ACS symposium series 1193* (pp. 29–42). Washington, DC: Oxford University Press. https://doi.org/10.1021/bk-2015-1193.ch003

Process oriented guided inquiry learning. Retrieved September 6, 2018, from USA. http://www.pogil.org

*Rau, M. A., Kennedy, K., Oxtoby, L., Bollom, M., & Moore, J. W. (2017). Unpacking "active learning": A combination of flipped classroom and collaboration support is more effective but collaboration support alone is not. *Journal of Chemical Education*, *94*, 1406–1414. https://doi.org/10.1021/acs.jchemed.7b00240

*Robert, J., Lewis, S. E., Oueini, R., & Mapugay, A. (2016). Coordinated implementation and evaluation of flipped classes and peer-led team learning in general chemistry. *Journal of Chemical Education*, *93*(12), 1993–1998. https://doi.org/10.1021/acs.Jchemed.6b00395

Ruiz-Primo, M. A., Briggs, D., Iverson, H., Talbot, R., & Shepard, L. A. (2011). Impact of undergraduate science course innovations on learning. *Science*, *331*, 1269–1270. https://doi.org/10.1126/science.1198976

*Ryan, M. D., & Reid, S. A. (2016). Impact of the flipped classroom on student performance and retention: A parallel controlled study in general chemistry. *Journal of Chemical Education*, *93*, 13–23. https://doi.org/10.1021/acs.jchemed.5b00717

*Saleh, T. A. (2011). Statistical analysis of cooperative strategy compared with individualistic strategy: An application study. *The Journal of Effective Teaching*, *11*, 19–27.

Seery, M. K. (2015). Flipped learning in higher education chemistry: Emerging trends and potential directions. *Chemistry Education Research and Practice*, *16*, 758–768. https://doi.org/10.1039/c5rp00136f

*Şen, Ş., & Yilmaz, A. (2016). The effect of process oriented guided inquiry learning (POGIL) on 11th graders' conceptual understanding of electrochemistry. *Asia-Pacific Forum on Science Learning and Teaching*, *17*, 5.

*Shachar, H., & Fischer, S. (2004). Cooperative learning and the achievement of motivation and perceptions of students in 11th grade chemistry classes. *Learning and Instruction*, *14*, 69–87. https://doi.org/10.1016/j.learninstruc.2003.10.003

*Shatila, A. (2007). *Assessing the impact of integrating POGIL in elementary organic chemistry*. (Doctoral dissertation): University of Southern Mississippi. Retrieved from ProQuest Dissertations & Theses database. (3289750)

*Shields, S. P., Hogrebe, M. C., Spees, W. M., Handlin, L. B., Noelken, G. P., Riley, J. M., & Frey, R. F. (2012). A transition program for underprepared students in general chemistry: Diagnosis, implementation, and evaluation. *Journal of Chemical Education*, *89*, 995–1000. https://doi.org/10.1021/ed100410j

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. https://doi.org/10.1037/a0033242

*Sisovic, D., & Bojovic, S. (2000). Approaching the concepts of acids and bases by cooperative learning. *Chemistry Education Research and Practice*, *1*, 263–275. https://doi.org/10.1039/a9rp90027f

*Sisovic, D., & Snezana, B. (2001). The elaboration of the salt hydrolysis concept by cooperative learning. *Journal of Science Education; Bogotá*, *2*, 19–23.

*Smetana, L. K., & Bell, R. L. (2014). Which setting to choose: Comparison of whole-class vs. small-group computer simulation use. *Journal of Science Education and Technology*, *23*, 481–495. https://doi.org/10.1007/s10956-013-9479-z

Stains, M., & Vickrey, T. (2017). Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices. *CBE—Life Sciences Education*, *16*, rm1. https://doi.org/10.1187/cbe.16-03-0113

*Stockwell, B. R., Stockwell, M. S., & Jiang, E. (2017). Group problem solving in class improves undergraduate learning. *ACS Central Science*, *3*, 614–620. https://doi.org/10.1021/acscentsci.7b00133

*Stoica, D., Chiru, L., & Chiru, C. (2012). Opportunity assessment for the introduction of the integrated learning unit in chemistry education. *Procedia—Social and Behavioral Sciences*, *46*, 4056–4060. https://doi.org/10.1016/j.sbspro.2012.06.196

*Straumanis, A., & Simons, E. A. (2008). A multi-institutional assessment of the use of POGIL in organic chemistry. In R. S. Moog & J. N. Spencer (Eds.), *Process oriented guided inquiry learning (POGIL). ACS symposium series 994* (pp. 226–239). Washington, DC: Oxford University Press.

*Talanquer, V., & Pollard, J. (2017). Reforming a large foundational course: Successes and challenges. *Journal of Chemical Education*, *94*, 1844–1851. https://doi.org/10.1021/acs.jchemed.7b00397

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology*, *2*, 85–112.

*Tarhan, L., & Acar, B. (2007). Problem-based learning in an eleventh grade chemistry class: 'Factors affecting cell potential'. *Research in Science & Technological Education*, *25*, 351–369. https://doi.org/10.1080/02635140701535299

*Tarhan, L., & Acar Sesen, B. (2012). Jigsaw cooperative learning: Acid–base theories. *Chemistry Education Research and Practice*, *13*, 307–313. https://doi.org/10.1039/c2rp90004a

*Tarhan, L., & Acar-Sesen, B. (2013). Problem based learning in acids and bases: Learning achievements and students' beliefs. *Journal of Baltic Science Education*, *12*, 565–578.

*Tarhan, L., Ayar-Kayali, H., Urek, R. O., & Acar, B. (2008). Problem-based learning in 9th grade chemistry class: 'Intermolecular forces'. *Research in Science Education*, *38*, 285–300. https://doi.org/10.1007/s11165-007-9050-0

*Tarhan, L., Ayyıldız, Y., Ogunc, A., & Sesen, B. A. (2013). A jigsaw cooperative learning application in elementary science and technology lessons: Physical and chemical changes. *Research in Science & Technological Education*, *31*, 184–203. https://doi.org/10.1080/02635143.2013.811404

*Tarhan, L., & Sesen, B. A. (2010). Investigation the effectiveness of laboratory works related to "acids and bases" on learning achievements and attitudes toward laboratory. *Procedia—Social and Behavioral Sciences*, *2*, 2631–2636. https://doi.org/10.1016/j.sbspro.2010.03.385

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*, 2113–2126. https://doi.org/10.1002/sim.1461

*Tien, L. T., Roth, V., & Kampmeier, J. A. (2002). Implementation of a peer-led team learning instructional approach in an undergraduate organic chemistry course. *Journal of Research in Science Teaching*, *39*, 606–632. https://doi.org/10.1002/tea.10038

Tien, L. T., Roth, V., & Kampmeier, J. A. (2004). A course to prepare peer leaders to implement a student-assisted learning method. *Journal of Chemical Education*, *81*, 1313. https://doi.org/10.1021/ed081p1313

*Tosun, C., & Taskesenligil, Y. (2013). The effect of problem-based learning on undergraduate students' learning about solutions and their physical properties and scientific processing skills. *Chemistry Education Research and Practice*, *14*, 36–50. https://doi.org/10.1039/c2rp20060k

*Turaçoğlu, İ., Alpat, Ş, & Ellez, A. M. (2013). Effects of jigsaw on teaching chemical nomenclature. *Education & Science/Egitim Ve Bilim*, *38*, 256–272.

*Üce, M., & Ateş, İ. (2016). Problem-based learning method: Secondary education 10th grade chemistry course mixtures topic. *Journal of Education and Training Studies*, *4*, 30–35. https://doi.org/10.11114/jets.v4i12.1939

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Vo, H. M., Zhu, C., & Diep, N. A. (2017). The effect of blended learning on student performance at course-level in higher education: A meta-analysis. *Studies in Educational Evaluation*, *53*, 17–28. https://doi.org/10.1016/j.stueduc.2017.01.002

Walker, L., & Warfa, A.-R. M. (2017). Process oriented guided inquiry learning (POGIL®) marginally effects student achievement measures but substantially increases the odds of passing a course. *PLoS One*, *12*, e0186203. https://doi.org/10.1371/journal.pone.0186203

*Wan Yahaya, W. A. J., & Tan, H. L. (2017). The effects of problem-based learning strategies and learning style on students' achievement and retention in a social network environment. *Presented at the International Technology, Education and Development Conference, Valencia, Spain*; 5237–5246. https://doi.org/10.21125/inted.2017.1225

Warfa, A.-R. M. (2016). Using cooperative learning to teach chemistry: A meta-analytic review. *Journal of Chemical Education*, *93*, 248–255. https://doi.org/10.1021/acs.jchemed.5b00608

*Webster, A. A., & Riggs, R. M. (2006). A quantitative assessment of a medicinal chemistry problem-based learning sequence. *American Journal of Pharmaceutical Education*, *70*, 89. https://doi.org/10.5688/aj700489

Wilson, S. B., & Varma-Nelson, P. (2016). Small groups, significant impact: A review of peer-led team learning research with implications for STEM education researchers and faculty. *Journal of Chemical Education*, *93*, 1686–1702. https://doi.org/10.1021/acs.jchemed.5b00862

*Yalçınkaya, E., Taştan-Kırık, Ö., Boz, Y., & Yıldıran, D. (2012). Is case-based learning an effective teaching strategy to challenge students' alternative conceptions regarding chemical kinetics? *Research in Science & Technological Education*, *30*, 151–172. https://doi.org/10.1080/02635143.2012.698605

*Yestrebsky, C. L. (2016). Direct comparison of flipping in the large lecture environment. In J. L. Muzyka & C. S. Luker (Eds.), *The flipped classroom volume 2: Results from practice ACS symposium series 1228* (pp. 1–18). Washington, DC: Oxford University Press. https://doi.org/10.1021/bk-2016-1228.ch001

*Yoruk, A. (2016). Effect of jigsaw method on students' chemistry laboratory achievement. *International Journal of Educational Sciences*, *15*, 377–381. https://doi.org/10.1080/09751122.2016.11890547

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Rahman T, Lewis SE. Evaluating the evidence base for evidence-based instructional practices in chemistry through meta-analysis. *J Res Sci Teach*. 2019;1–29. https://doi.org/10.1002/tea.21610